AWS Inferentia vs AWS Trainium2

Side by side datacenter GPU specifications from the Flopper.io database.

OVERVIEW

	AWS Inferentia	AWS Trainium2
FP32 TFLOPS	–	**181.0**
Memory	**8 GB**	**96 GB**
Memory Bandwidth	**50 GB/s**	**2.9 TB/s**
TDP	–	–

PERFORMANCE BY PRECISION

Precision	AWS Inferentia peak TFLOPS (dense)	AWS Trainium2 peak TFLOPS (dense)
FP64	–	–
FP32	–	181.0
TF32	–	667.0
BF16	64.0	667.0
FP16	64.0	667.0
FP8	–	1299.0
FP6	–	–
FP4	–	–
INT8	128.0	–

SPECIFICATIONS

	AWS Inferentia	AWS Trainium2
Architecture	NeuronCore-v1	NeuronCore-v3
Process Node	–	–
Launch Year	2018	2024
Form Factor	–	–
Memory Type	DDR4	–
Interface Width	–	–
Interconnect	NeuronLink-v1	NeuronLink-v3

	AWS Inferentia	AWS Trainium2
TDP	–	–
FP32 Efficiency	–	–